Journal of Experimental Psychology:
Learning, Memory, and Cognition
2002, Vol. 28, No. 3, 458–467


0278-7393/02/$5.00 DOI: 10.1037//0278-7393.28.3.458

# Statistical Learning of Higher-Order Temporal Structure From Visual Shape Sequences

József Fiser and Richard N. Aslin
University of Rochester

In 3 experiments, the authors investigated the ability of observers to extract the probabilities of successive shape co-occurrences during passive viewing. Participants became sensitive to several temporal-order statistics, both rapidly and with no overt task or explicit instructions. Sequences of shapes presented during familiarization were distinguished from novel sequences of familiar shapes, as well as from shape sequences that were seen during familiarization but less frequently than other shape sequences, demonstrating at least the extraction of joint probabilities of 2 consecutive shapes. When joint probabilities did not differ, another higher-order statistic (conditional probability) was automatically computed, thereby allowing participants to predict the temporal order of shapes. Results of a single-shape test documented that lower-order statistics were retained during the extraction of higher-order statistics. These results suggest that observers automatically extract multiple statistics of temporal events that are suitable for efficient associative learning of new temporal features.

Our visual experience consists almost entirely of spatiotemporal events created by observer movement through the visual array (through eye, head, and body movements) and/or by independent movements of objects with respect to the static environment. Therefore, the computational task facing the visual system during the interpretation or learning of new visual scenes is one of extracting spatiotemporal correlations. A simple classification of these spatiotemporal correlations is illustrated in Figure 1, in which low-level visual analyzers respond to the presence of one or more object features in two or more time frames. Several different types of spatiotemporal correlations could be present across multiple time frames: (a) no change in the features or the position of the object, (b) a change in object position, with no change in the features of the object, (c) a smooth transformation of one or more features, with no change in position, or (d) an abrupt change in one or more features of the object, with or without a change in position. Type A is a prototypical example of static vision, in which high spatial correlations among a set of features are present for extended periods of time. Type B is an example of object motion (short- or long-range, depending on the magnitude of the change in position across frames). Type C is an example of object rotation (in either 2-D or 3-D) or object deformation. And Type D is an example of a sudden replacement of one object with a new object, or a saccade to a new target.

This classification highlights the fact that sensitivity to temporal statistics across a variety of different scales plays a crucial role in most aspects of vision, including extracting temporal correlations between small patches of two sequential images falling onto the retina during continuous viewing, integrating visual information across saccades, or identifying visual event sequences. Indeed, the two-dimensional spatial nature of visual input (at the level of the retinal image) has diverted attention from the equally important fact that visual information is also defined in the temporal domain, just like auditory information. A case can be made that there is no visual information without temporal information, and thus decoding the processing mechanisms of temporal correlations is not just related to some special cases or nonvisual sensory modalities, but is an essential requirement for understanding vision.

It is important to note that in Type A, B, and C tasks illustrated in Figure 1, the spatiotemporal correlations are computed from highly redundant images, where the analyzers operate over brief temporal intervals. As a result, even in cluttered scenes, the correspondence between object features across time frames is relatively straightforward. And, in fact, the visual system has evolved a number of low- and mid-level analyzers that rapidly and efficiently extract spatiotemporal correlations in Type A tasks (Chubb, Econopouly, & Landy, 1994; Julesz, 1981), Type B tasks (Watamaniuk & Sekuler, 1992; see also Lee & Blake, 1999), and Type C tasks (Johansson, 1973). In contrast, Type D tasks involve low spatial redundancy over frames (typically different images), suggesting that feature matching across time frames may not provide useful information for computing spatiotemporal correlations. In the absence of any feature matches across time frames, the task becomes one of extracting the serial order of a set of discrete objects encoded in memory. Therefore, Type D tasks are ideally suited for studying temporal correlations with the least inference from the automatic activation of low- and mid-level visual analyzers, which are specialized for detecting spatial correlations.

József Fiser and Richard N. Aslin, Department of Brain and Cognitive Sciences and Center for Visual Science, University of Rochester.

This research was supported by James S. McDonnell Foundation Postdoctoral Fellowship 96-32 awarded to József Fiser and by National Science Foundation Research Grant BCS-9873477 awarded to Richard N. Aslin. We are grateful to Elizabeth Gramzow, Conni Augustine, Rebecca DaMore, and Rachel Heafitz for testing the participants, and to Elissa Newport and Ruskin Hunt for helpful comments on an earlier version of this article.

Correspondence concerning this article should be addressed to Richard N. Aslin, Department of Brain and Cognitive Sciences, Meliora Hall, University of Rochester, Rochester, New York 14627-0268. E-mail: aslin@cvs.rochester.edu

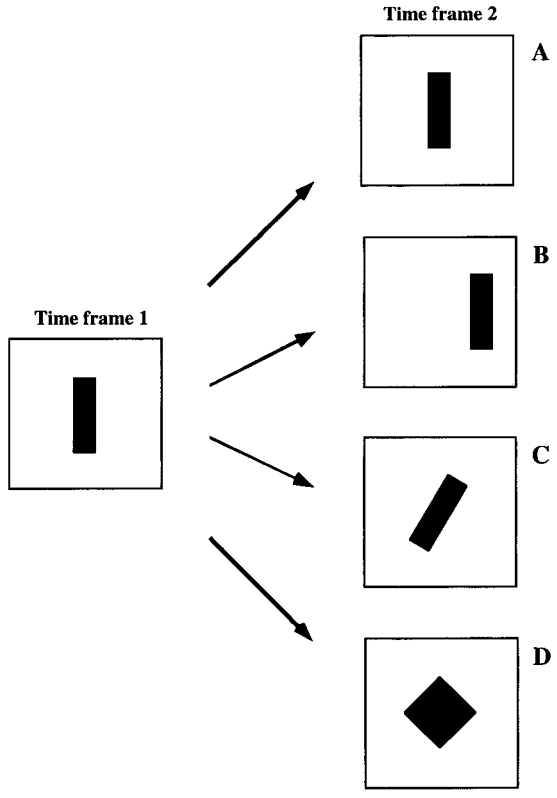


*Figure 1.* Illustration of four types of spatiotemporal correlations across a two-image sequence: (A) no change (identity), (B) object movement, (C) smooth feature transformation, and (D) object change.

This type of task, which is the focus of the present series of experiments, enables us to study *which* temporal statistics are used to temporally bind discrete objects across scenes.

In contrast to the visual domain, in which spatial correlations are so fundamental to object perception and relatively little attention has been directed to temporal correlations, the auditory domain involves stimuli that are fundamentally temporal in nature. For example, speech and music are complex time-varying modulations of frequency and intensity whose primitives (phonemes and notes) are concatenated into a temporal stream to form higher level groupings (syllables/words and melodies). The importance of temporal correlations in the auditory domain has been demonstrated by Saffran, Newport, and Aslin (1996), who showed that adult listeners are remarkably sensitive to auditory order information in speech streams and can rapidly learn such order information from mere exposure to initially unfamiliar stimuli. Subsequent experiments demonstrated this same temporal-order sensitivity in children (Saffran, Newport, Aslin, Tunick, & Barrueco, 1997) and in 8-month-old infants (Aslin, Saffran, & Newport, 1998; Saffran, Aslin, & Newport, 1996), as well as in a nonhuman primate (Hauser, Newport, & Aslin, 2001). These studies provide strong evidence that a robust learning mechanism is available in adults, children, infants, and monkeys to extract the sequential structure of rapid auditory events. Saffran, Johnson, Aslin, and Newport (1999) also demonstrated that these abilities are not unique to speech, because adults and infants show the same segmentation and grouping of rapid streams of tones.

Fiser and Aslin (2001) extended these results from the auditory temporal domain to the visual spatial domain. They created a large ($> 100$) set of six-element scenes, which adults viewed at a rate of one every 2 s, with no instructions provided to guide their learning. After this passive familiarization to the scenes, participants judged in a two-alternative forced-choice posttest which of two scenes was more familiar. One of the two test scenes contained a subset of elements in a spatial configuration identical to those presented in some of the familiarization scenes, and the other test scene was composed of the same number of elements but in a novel or less predictable spatial configuration. Participants reliably chose the familiar spatial configuration over the novel or less predictable spatial configuration, even though there were no instructions provided during familiarization, and no feedback given during familiarization or test. These results show that spatial correlations can be extracted from a large set of multielement scenes based solely on the statistics of the spatial configuration of the elements. What remains unclear is whether these statistical learning abilities are specialized for the dominant dimensions of different sensory modalities—temporal in the auditory domain, where rapid temporal events are ubiquitous, and spatial in the visual domain, where spatial relations are of paramount importance—or whether they also operate in the temporal dimension for the visual domain.

Several recent results suggest considerable variability in the learning of temporal correlations across scenes. For example, change-blindness tasks (Simons & Levin, 1997), the integration of information across saccadic eye movements (Henderson, 1997; Irwin, 1996), and memory of distractor items in a visual search task (Horowitz & Wolfe, 1998) all suggest minimal access to temporal correlations. Other results suggest that context (e.g., preceding scenes in the sequential presentation of many scenes) implicitly promotes associations across scenes (Chun & Jiang, 1998, 1999; Chun & Nakayama, 2000; Olson & Chun, 2001), and that target priming can facilitate visual search (Maljkovic & Nakayama, 1994, 1996, 2000). Moreover, there is an extensive literature that demonstrates learning of new associations across scenes in humans (Cleeremans & McClelland, 1991; Cohen, Ivry, & Keele, 1990; Reed & Johnson, 1994; Stadler, 1992) and infants (Kellman & Short, 1987). Thus, under some circumstances, there is clear evidence that temporal correlations can be learned across scenes, even when no feedback is provided to guide the learning task. Therefore, the question addressed by the present study was not whether it can be done, but rather *how* and *what kind of* temporal correlations can be used to develop temporal features of events, and whether the rules used to extract these temporal correlations are similar to those for spatial features under static conditions.

The present series of experiments investigated these issues by extending the paradigm used by Saffran, Aslin, and Newport (1996) from the auditory to the visual modality. Three questions were the focus of our investigations: (a) Is there evidence for unsupervised statistical learning of the temporal order of visual stimuli in human adults in the absence of first-order temporal statistical signatures?; (b) Does such learning involve concurrent computations of both first- and higher-order temporal statistics?; and (c) In what form is the information extracted about temporal statistics applied in recognition tasks?

Both the spatial and the temporal statistics in our experiments were quite simple, thereby enabling us to precisely control all

relevant statistical relations and to avoid the "combinatorial explosion" problem that occurs with more complex stimulus sets. For the temporal statistics, we used changes in a single shape that moved back and forth across a visual display. For the spatial statistics, we used a small set of highly discriminable simple shapes. The underlying assumption of this approach was that, on the basis of low-level visual analyzers, participants could easily identify the shapes as they were presented. Thus, any difficulty they had in extracting the temporal structure of the stream of shapes could not be attributed to the difficulty of individual shape discrimination.

## Experiment 1

The goal of the first experiment was to determine whether human adults are sensitive to the temporal structure of a continuous sequence of shapes when the frequency of the individual shapes in the sequence was equated. Twelve shapes were organized into four temporal triplets, so that if one element of the triplet appeared on the screen, the next shape in the triplet always appeared next in the sequence. The shapes were presented in temporal succession at a uniform rate so that duration could not be used to group or segment the shape sequence. Because the frequency of individual shapes and the frequency of shape triplets were uniform across the shape sequence, the only structure the participants could rely on for grouping and segmentation was the ordering of the shapes. Thus, successful learning minimally required the extraction of temporal-order statistics among the pairs or triplets of shapes. Moreover, because participants were not informed that the shapes were organized into triplets, or that they should attend to any temporal grouping of the shape sequences, the task involved statistical learning that was unsupervised.

### *Method*

*Participants.* Undergraduates from the University of Rochester, who were paid $6 per session for their participation, served as the participants in each of the experiments reported in this study. All participants were naive with respect to the purpose of the experiment and participated only in one experiment. In Experiment 1, there were 8 participants.

*Stimuli.* Twelve simple, arbitrary, black shapes were generated on a white background (see Figure 2). The largest extent of the individual shapes was equated at 5.65°. Special care was taken to provide distinct shapes so that even after considerable low-pass filtering they were not easily confused with each other.

A continuous movie of a shape sequence was generated using Macromedia Director (Version 8.0, Macromedia, San Francisco; see Figure 3).

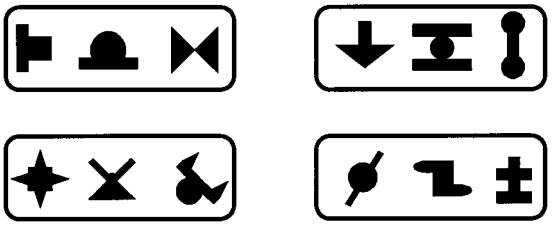

*Figure 2.* The 12 basic shapes, grouped into four triplets, that were used in all of the experiments. The enclosing boxes are only for demonstrative purposes.

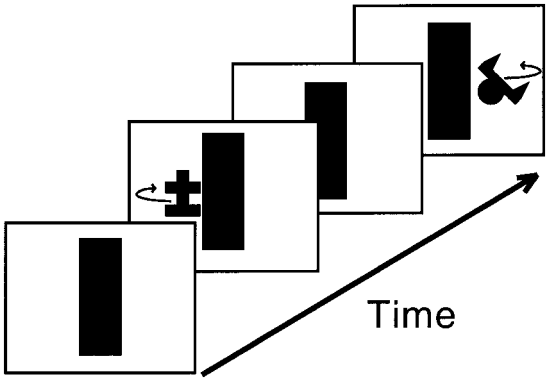


*Figure 3.* Sample frames from the video animation used to present the continuous stream of shape sequences. A single shape moved in a continuous movement from behind the vertical occluder on either the left or right side, moved to the edge of the screen, moved back behind the occluder, and then changed to a different shape when it reappeared on the other side of the occluder. (Arrows demonstrate the motion of shapes and were not visible in the experiments.)

A 5.7°-wide × 15.4°-long static, black, vertical bar was positioned in the middle of a 21-in. 1,024-pixel × 768-pixel Sony computer monitor for the entire duration of the movie. The movie consisted of a single shape that moved smoothly with constant speed, horizontally back and forth between the sides of the screen, halfway between the top and bottom of the screen. As the moving shape came into contact with the vertical bar, it was gradually occluded by the bar. When the shape was completely occluded, it changed to 1 of the other 11 shapes. The changed shape continued on the same trajectory with no interruption, gradually emerging from behind the vertical bar on the other side of the screen. It took exactly 1 s for a shape starting from the center (covered completely by the vertical bar) to move to the side of the screen and return to the initial position behind the bar.

The change from one shape to the other was not random but rather followed a strictly imposed structure. Each shape appeared an equal number of times during the shape sequence. The 12 shapes were grouped arbitrarily into four *base triplets*, shown by the three adjacent shapes in Figure 2, and referred to as *A-B-C*, *D-E-F*, *G-H-I*, and *J-K-L*. A base triplet represented a consistent structure in the stream; that is, if *A-B-C* was a base triplet, then whenever *A* appeared on the screen, it always changed next to *B*, which in turn always changed to *C*, when passing behind the bar. After *C*, only the first shape element from one of the other three base triplets could be presented (i.e., immediate repetition of a base triplet was not allowed). Each of the remaining base triplets followed a given base triplet with equal probability. This structure created a continuous flow of shapes in which the four base triplets followed each other in random order (subject to the constraints described above), and assured that each individual element appeared on both sides of the central occluding bar equally often.

The underlying structure of this shape sequence consisted of the following temporal statistics. The first-order statistics, the probability of appearance of each of the 12 shapes, was equated and therefore provided no information for grouping or segmentation. The probability of any shape—for example, $P(A)$ or $P(L)$—was .083. The second-order statistics, defined by the joint probability of pairs of successive shapes, varied by position within the shape sequence as follows. The joint probability of shape pairs within any of the four base triplets—for example, $P(A,B)$ or $P(K,L)$—was .083, whereas the joint probability of shape pairs spanning two base

triplets—for example, $P(C,D)$ or $P(I,A)$—was .027.[1] The third-order statistics of adjacent triplets were redundant with the second-order statistics. That is, the joint probability of any base triplet—for example, $P(A,B,C)$—was .083, and the joint probability of any triplet spanning a base triplet boundary—for example, $P(C,D,E)$ or $P(H,I,A)$—was .027.

*Procedure.* A 6-min movie was generated by concatenating 96 base triplets in a semirandom fashion so that no repetitions of base triplets or triplet pairs (e.g., $Triplet_1$, $Triplet_3$, $Triplet_1$, $Triplet_3$) occurred, and the number of base triplets was identical in each third of the movie. Participants were instructed to view the movie so that they would be able to answer questions about what they saw. No information was given to the participants about the potential temporal sequencing of the shapes. After the movie, a two-interval forced-choice (2IFC) test was given to the participants, with 32 test pairs. Each of these test pairs consisted of one of the four base triplets and one of four "impossible" triplets that were generated from three successive shapes that had never occurred after each other in the training shape sequence. That is, the probability for each of the two shape pairs in the impossible triplets appearing in the movie during training—for example, $P(B,E)$—was 0. Each base triplet was paired with each impossible triplet in two different orders to yield the 32 test pairs. In each test pair, the two triplets were shown in a left–right–left sequence, with a 1-s pause between triplets. Participants were instructed to judge which of the two test triplets was more familiar on the basis of the training movie in a balanced 2IFC posttest. They were allowed unlimited time to make their decision, but typically responded within 2–3 s.

### *Results and Discussion*

All 8 participants easily discriminated the base triplets from the impossible triplets (mean percent correct = 95%), $t(7) = 25.57$, $p < .0001$, despite the fact that all 12 shapes appeared with equal probability, that the four impossible triplets contained the same shapes as the base triplets, and that there were no segmentation cues (other than the sequential statistics) in the familiarization sequence. Thus, participants must have based their judgments on the rapid learning from the shape sequence of a higher-order temporal statistic than the first-order statistic of single-shape appearance probability. The simplest of these statistics is the joint probability of two shapes appearing in a particular order in the sequence. Of most importance, the participants became sensitive to this temporal-order information in an unsupervised manner, without any instruction to pay attention to pairs or triplets in the shape sequence.

## Experiment 2

The first experiment demonstrated that adults are naturally sensitive to some higher-order temporal statistics, minimally to the joint probability of shape pairs from a continuous stream of shapes. However, the test items used in Experiment 1 were maximally different in their joint probabilities. Both of the shape pairs in the base triplets had joint probabilities of .083, whereas both of the shape pairs in the impossible triplets had joint probabilities of 0. In Experiment 2, we asked two further questions about the types of higher-order temporal statistics that adults can compute from visual shape sequences. First, can participants learn temporal statistics from shape sequences when the joint probabilities of shape pairs in the test items are all nonzero? Second, do participants rely on the joint probabilities of all three shapes in a base triplet, or do they rely on information about shape pairs when judging the familiarity of test triplets from the shape sequence? This second question has implications for the manner in which statistical information extracted from the shape sequence is represented and used.

### *Method and Stimuli*

The shape stimuli and the training movie (and hence the imposed temporal structure of the movie) were exactly the same as in Experiment 1. The difficulty of the 2IFC discrimination task was increased by replacing the impossible triplets with *part triplets*. Part triplets consisted of either the last shape of one base triplet and the first two shapes of another base triplet (3-1-2) or the last two shapes of one base triplet and the first shape of another base triplet (2-3-1).[2] Thus, in contrast to an impossible triplet, which never appeared in the training sequence, a part triplet did appear in the training sequence, but with lower probability. The joint probability of both types of part triplets—for example, $P(C,D,E)$ and $P(B,C,D)$—was .027, a value one third that of the base triplets. However, the order of joint probabilities of shape pairs within the two types of part triplets differed. For the 3-1-2 part triplets, the first shape pair had a joint probability of .027 and the second shape pair had a joint probability of .083. Thus, the only difference between this part triplet and a base triplet was the joint probability of the first shape pair. In contrast, for the 2-3-1 part triplets, the only difference between the base triplets and the part triplets was in the second shape pair. The number of different part triplets in the training movie was balanced. If participants relied solely on the first shape pair during the 2IFC posttest, then the 3-1-2 part triplets but not the 2-3-1 part triplets should be discriminable from the base triplets. However, if participants used both shape pairs that defined a base triplet equally, performance should be equal in the two different part-triplet test conditions. Separate groups of 8 participants were tested in the two part-triplet conditions.

### *Results and Discussion*

The results of the two part-triplet tests along with the results from Experiment 1 are shown in Figure 4. The results from the 3-1-2 condition revealed that participants reliably discriminated the base triplets from the part triplets ($M = 69.1\%$), $t(7) = 4.75$, $p < .0021$. Thus, even when the only difference between the test triplets was the nonzero joint probability of one of the two shape pairs, grouping and segmentation from the shape sequence was possible. This demonstrates that learning in the visual domain can be accomplished when the differences in magnitude between higher-order temporal statistics are quite subtle.

However, the results from the 2-3-1 condition indicated that when the order of the joint probabilities of the shape pairs was reversed in the part triplets, performance dropped to chance ($M = 55.4\%$), $t(7) = 1.14$, $p > .29$. This drop was significant compared with the results in the 3-1-2 condition, $t(14) = 2.18$, $p < .05$. It should be assumed that the statistical information available in the 3-1-2 and the 2-3-1 tests was exactly the same: A base triplet had to be discriminated from a triplet having one pair with

[1] The definition of the joint probability of a shape pair $P(A,B)$ in our task is the probability that a random selection of two successive shapes in the stream will result in the ordered shape pair *A,B*. Similarly, the definition of the joint probability of a shape triplet $P(A,B,C)$ is the probability that a random selection of three successive shapes in the stream will result in the ordered shape triplet *A,B,C*.

[2] For example, some possible 3-1-2 part triplets were *C-D-E, F-G-H*, or *C-G-H*. Some of the possible 2-3-1 part triplets were *B-C-D, H-I-A*, or *E-F-J*.

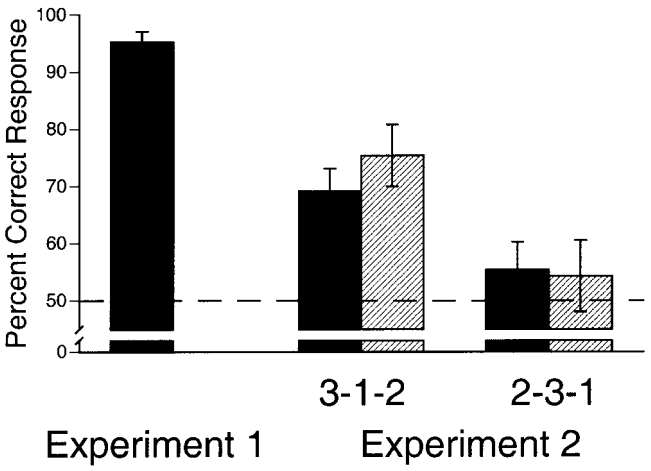


*Figure 4.* The results of Experiments 1 and 2. The *y*-axis is truncated below 50%, which is chance performance in both experiments. Performance in Experiment 1 (leftmost solid bar) and on the 3-1-2 test of Experiment 2 was significantly above chance, whereas performance on the 2-3-1 task was not (solid bars). The striped bars indicate performance in Experiment 2A in which the training was extended from 6 to 12 min. More extensive training did not eliminate the difference between the results of the 3-1-2 and the 2-3-1 type of experiments, and it did not elevate the results of the 2-3-1 experiment to significantly above chance performance. Error bars show standard errors.

joint probability of $P = .083$ and a second pair with $P = .027$. Thus, there appears to be a constraint on the statistical information that is used to make temporal-order judgments. Participants appeared to base their decision on the first available higher-order statistic contained in the test triplets: the joint probability of the first shape pair. In the 3-1-2 condition, but not in the 2-3-1 condition, the joint probability of the first shape pair differed from the base triplet. Of interest, this same asymmetry was obtained by Saffran, Newport, and Aslin (1996) in the auditory domain, although the structure of their part triplets was slightly different (X-1-2 and 2-3-X, where *X* was an "out-of-order" syllable from the speech stream).

### *Experiment 2A*

The difference in performance on the two test conditions in Experiment 2 could simply reflect the greater difficulty of attending to the statistics of the second shape pair in the test triplets. To explore this alternative, we reran both testing conditions on two new groups of 8 participants with a training session that was twice as long as in Experiment 2 (12 min vs. 6 min). Participants' performance in the 3-1-2 condition was again significantly above chance ($M = 75.4\%$), $t(7) = 4.71$, $p < .0022$, improving slightly but not significantly as compared with Experiment 2. Performance in the 2-3-1 condition remained unchanged from Experiment 2 and was not significantly above chance ($M = 54.3\%$), $t(7) = 0.69$, $p > .51$. The difference in performance between the two conditions remained significant, $t(14) = 2.56$, $p < .05$. Thus the extended familiarization did not change the pattern of results from Experiment 2, suggesting that participants indeed treated the two test conditions differently.

## Experiment 3

The previous experiments established that adults are sensitive to at least one higher-order temporal statistic of a shape sequence, which we characterized by the joint probabilities of two or more simple single shapes appearing in a particular consecutive order. However, there was another statistic present in these shape sequences that may have been used by the participants to learn the underlying temporal-order information. That other statistic is the conditional probability of successive shape pairs, as given by the following formula:

$$P(Y|X) = P(X, Y)/P(X),$$

where $P(X)$ is the probability of Event X, $P(X,Y)$ is the joint probability of the sequence of Events X and Y, and $P(Y|X)$ is the conditional probability of Event Y given Event X. In the present context, $P(X)$ is the probability that a simple shape appears in a given position in the movie sequence, $P(X,Y)$ is the joint probability that the ordered shape pair *X–Y* appears consecutively in the movie, whereas $P(Y|X)$ is the conditional probability that Shape Y follows Shape X given that *X* appeared. The goal of the present experiment was to determine whether human observers are sensitive to conditional probabilities when the joint probabilities have been equated.

Given that participants in Experiments 1 and 2 could extract information about shape sequences using joint probabilities, why is it interesting to determine whether they are also able to compute conditional probabilities? The fundamental difference between joint and conditional probabilities of two events is that the first signals the probability of co-occurrence, whereas the latter measures the predictive power of one event with respect to another. One should consider the following simple example in which many different events occur in a long sequence. Three of these events—*A*, *B*, and *C*—occur with unequal frequency, such that *A* is more frequent than *B* or *C*—that is, $P(A) \gg P(B), P(C)$. It should be assumed that after Event A, many other events can follow, among them one seventh of the time, Event B. In contrast, assume that Event C is followed almost exclusively by Event B. If the frequency of Event C—that is, $P(C)$—is one seventh of the frequency of Event A, then the following situation occurs. The joint probabilities of $P(A,B)$ and $P(C,B)$ are about equal; in other words, *B* following *A* happens as frequently as does *B* following *C*. However, when *A* appears in a sequence, we have very little certainty (one seventh to be exact) that the next event will be *B*, whereas the appearance of *C* almost completely assures us that the next event will be *B*. In other words, paying attention to the joint probabilities would rank the events *A*-followed-by-*B* and *C*-followed-by-*B* as equally important, yet *C* is superior to *A* in terms of reducing our uncertainty (enhancing the predictability) of future events.

The importance of detecting conditional probabilities comes from a long line of research based on information theoretical considerations. This research posits that reducing uncertainty, or increasing the efficiency of the coding of sensory information, is essential for associative learning and ultimately for normal functioning in the brain (Atick, 1992; Attneave, 1954; Barlow 1961, 1989; von der Malsburg, 1981). These theories rely heavily on the implicit assumption that humans are sensitive to and use conditional probabilities embedded in sensory inputs, because otherwise

efficient learning and redundancy reduction would be implausible. Conditional probabilities have also been shown in animals to better characterize what they learn about associative contingencies than simple frequencies of occurrence or co-occurrence (Rescorla & Wagner, 1972; see also Gallistel, 1990). Sensitivity to conditional probabilities has also been shown by human infants in the learning of speech sequences (Aslin et al., 1998), and by adults in the learning of spatial correlations in visual patterns (Fiser & Aslin, 2001) and spatiotemporal correlations in a serial reaction time (SRT) task (Hunt & Aslin, 2001).

In the structure imposed on the shape sequences in the previous experiments, each simple shape appeared the same number of times—that is, $P(X) = k$. As a result, joint probabilities and conditional probabilities were correlated. Thus, it is not possible from Experiments 1 and 2 to determine whether grouping and segmentation were based on joint or conditional probability computations from the shape sequences. To address this question, the structure of the shape sequences in the training movie of Experiment 3 was changed. As a result, in the test session it became possible to compare triplets that appeared exactly the same number of times during training—that is, $P(A,B,C) = k$—but for which the conditional probabilities were very different. Such a change in structure involved unequal frequencies of base triplets. This structure also enabled us to examine a second question, whether participants are also sensitive to the differences in the frequencies of individual shapes.

### *Method and Stimuli*

The 12 simple shapes and the method of presentation in Experiment 3 were identical to those of Experiments 1 and 2. However, the imposed structure of the shape sequences was changed by making the frequency of two of the four base triplets double that of the other two base triplets. As in Experiments 1 and 2, base triplets did not repeat in immediate succession. For counterbalancing reasons, the number of triplets in the training movie increased slightly from 96 to 108 triplets. In addition, the total duration of the movie was increased to 12 min because extraction of conditional probabilities, when the joint probabilities were equated, was predicted to be a more difficult task.

Because participants in the test phase of Experiment 2 based their judgments of familiarity on the initial shape pair of a triplet, the test phase of Experiment 3 assessed sensitivity to the initial shape pair of triplets and part triplets rather than to the entire triplets. After exposure to the movie, sensitivity to conditional probabilities was tested by comparing the initial shape pairs of the two infrequent base triplets with the initial shape pairs of the two part triplets formed from the two frequent base triplets. Because these part triplets consisted of the 3-1-2 pattern that was learned in Experiment 2, the first transition within the part triplet was the between-base-triplets transition (3-1). The sequencing constraints in the training movie resulted in the *tested* shape pairs from the base triplets appearing in the training movie exactly the same number of times as the 3-1 shape pairs from the part triplets; thus, the joint probabilities of the tested shape pairs from the base and part triplets were exactly the same. In contrast, there was a pronounced difference between the conditional probabilities of the first shape pair of the base triplets and the part triplets. In the base triplet, both shape pairs had conditional probabilities of 1.0 (Shape A perfectly predicted Shape B, and Shape B perfectly predicted Shape C). However, the conditional probability of the first shape pair of the part triplet was .5, because half of the time the last shape of a frequent base triplet was followed by one of the infrequent base triplets, and the other half of the time it was followed by the other frequent base triplet. This difference in the conditional probability of the first shape pair was the only statistical information that participants could use to discriminate between the tested shape pairs.

The test phase consisted of two parts. In the first part, sensitivity to the difference in conditional probabilities was tested with shape pairs as described above. Specifically, the sequence of the first two shapes of the rare base triplet, a *base pair*—$P(B|A) = 1.0$—was compared with the sequence of the last Element I of one of the frequent triplets followed by the first Element J of the other frequent triplet, a *part pair*—$P(J|I) = .5$. As in Experiments 1 and 2, participants had to choose the pair that looked more familiar on the basis of the familiarization movie. In the second part of the test, single shapes from the frequent triplets were compared with single shapes from the infrequent triplets. Participants were asked to choose the shape that appeared more frequently in the familiarization movie.

### *Results and Discussion*

Twenty participants were tested in the experiment. As shown by the dark bars in Figure 5, participants selected base pairs reliably more often than part pairs ($M = 66.3\%$), $t(19) = 2.73$, $p < .015$. This suggests that the difference between higher and lower conditional probabilities of shape pair transitions within the base and part triplets was detected and used by participants in the discrimination task. Participants' performance was also reliably above chance in selecting the more frequent single element in the second part of the test ($M = 85.8\%$), $t(19) = 11.38$, $p < .0001$. This suggests that the participants were able to encode both first-order and higher-order temporal statistics from the visual stream of shape sequences. Although we cannot conclude that participants in the preceding experiments based their performance on conditional probabilities (because joint probability statistics were available), the present experiment demonstrates that even when differences in joint probability statistics are not available, participants can use conditional probability statistics to group and segment a shape sequence.

### *Experiment 3A*

In Experiment 2A, we found that doubling the familiarization period did not change the pattern of results, leaving intact the significant difference between detecting part triplets of the 3-1-2 and 2-3-1 types. This can be explained by assuming that, after reaching a plateau, the effect of the additional familiarization offered only a small but nonsignificant advantage in extracting joint probabilities. However, in general, there must be a phase during the early extraction process when increased familiarization moves performance from chance to above-chance levels. To ensure that the results of Experiment 3 were not due to some idiosyncracies of the test items, independent of familiarization, we reran Experiment 3 with half the duration of familiarization.

To avoid confounding effects of less power, the number of participants was increased from 20 to 32. The results are shown by the hatched bars in Figure 5. With reduced familiarization, participants performed significantly worse on the pair test than in Experiment 3, $t(50) = 2.09$, $p < .05$, and more important, they could not reliably discriminate the base pairs from the part pairs ($M = 48.1\%$), $t(31) = 0.33$, $p > .73$. They could, however, discriminate the single shapes in the movie based on their frequency of occurrence ($M = 73.9\%$), $t(31) = 7.45$, $p < .0001$, although this performance was also significantly worse than par-

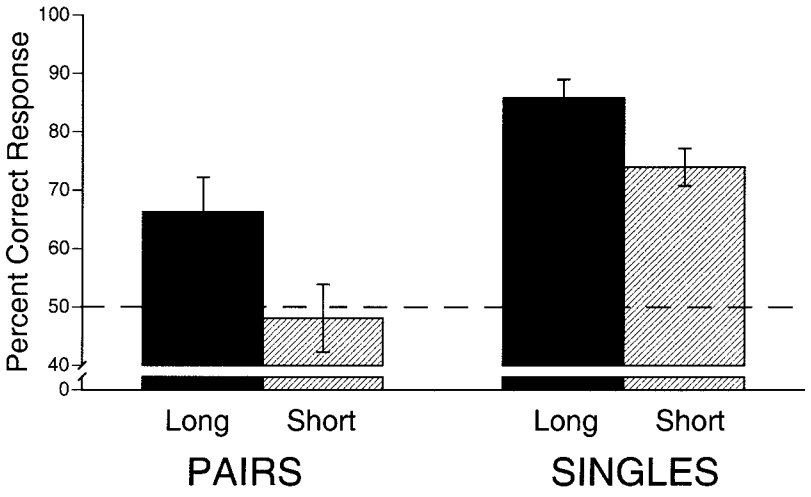


*Figure 5.* The results of Experiment 3. The bars on the left indicate the percent of selecting the frequency-balanced shape pair with 1.0 conditional probability as the more familiar shape pair in the pair two-interval forced-choice (2IFC) test over the shape pair with .5 conditional probability. The bars on the right indicate the percent of selecting the more frequent single shape correctly in the single 2IFC test. Solid bars are for 25 min of training and striped bars are for 12.5 min of training. Only the shorter training in the pair test failed to exceed chance performance. Error bars show standard errors.

ticipants' performance with single elements in Experiment 3, $t(50) = 2.49$, $p < .02$. Taken together, these results confirm that exposure time does matter in extracting descriptions of complex shape sequences. These results also suggest that participants required more exposure to extract higher-order statistical characteristics of sequences (i.e., conditional probabilities) than to extract lower-order statistics.

## General Discussion

In a series of experiments, we found that human adults are highly sensitive to the temporal statistics of image sequences. They learned these temporal-order statistics very rapidly, in as little as 6 min, and without an overt task or explicit instructions. Moreover, sequences of shapes presented during familiarization were distinguished from novel sequences of familiar shapes (Experiment 1), as well as from shape sequences that were seen during familiarization (part triplets) but less frequently than base triplets (Experiment 2). Because all 12 shapes were presented with the same frequency during familiarization in these studies, our results demonstrate that human observers are sensitive to at least the joint probability of successive shape pairs. Moreover, the results of Experiment 3 revealed that observers were sensitive to another statistic embedded in the shape sequences: the conditional probabilities of successive shape pairs. Even when joint probabilities did not differ, conditional probabilities were automatically computed, presumably in an attempt to predict the temporal order of shapes in the sequence. Sensitivity to this statistic improved with prolonged exposure during familiarization, but became significant only after sensitivity to first-order statistics emerged, suggesting that simpler statistics are extracted first. Despite this progression in extracting higher-order statistics, the results of the single-shape test documented that sensitivity to lower-order statistics was retained, suggesting that a range of statistics are available to apply to as yet unforeseen tasks.

An intriguing aspect of the results from Experiment 2 is that observers failed to discriminate between base triplets and part triplets when the initial shape pair of the part triplet was from within a base triplet. This failure cannot be attributed to an inability to discriminate certain shape pairs, because when the test phase involved pairs rather than triplets (Experiment 3) the within-triplet shape pairs were reliably discriminated from the between-triplet shape pairs. Therefore, the failure to use differences in joint probability that were present in the second shape pair of a part triplet is presumably due to the test context, indicating that when two triplets were being compared, the second shape pair in the triplets did not carry much weight in the discrimination judgment.

Does this difference in performance indicate that participants in the 3-1-2 group "learned" the triplets, whereas participants in the 2-3-1 group did not? Such an account is both highly implausible and unrelated to the primary importance of our findings. The implausibility stems from the fact that the two groups of participants performed differently despite identical exposure during the familiarization phase. Moreover, although the sequence of shapes during familiarization was constrained by its triplet structure, there was no nonstatistical information (e.g., pauses or duration differences) by which participants could group the shapes into triplets. Thus, from the point of view of the participants, they had no explicit knowledge of the triplet structure until the moment that they read the instructions for the posttest. Given the nature of our 2IFC posttest, it is unclear whether participants implicitly extracted triplets during the familiarization phase, or whether they simply became sensitive to the pairwise statistical relations present in the stream of shapes. Our claims center on learning temporal correlations across scenes, not on the learning of triplets per se. Thus, although identifying triplets can be cast into this framework,

pairwise shape information present during the familiarization phase is completely sufficient to discriminate between familiar and novel triplets. Therefore, the smallest step across time—learning correlations between two successive frames—is the mechanism by which we presume participants solved our posttest, and which accounts for the difference between the results with 3-1-2 and 2-3-1 triplets.[3]

Another intriguing question arising from our study is the order in which various statistics—for example, joint or conditional probabilities—are extracted as the stimulus materials are presented during familiarization.[4] From our results, it is clear that sensitivity to first-order correlations (appearance frequency) precedes both joint and conditional probabilities. However, once this low-order information is acquired, there is no a priori reason to favor either of these second-order statistics because each can be computed from the other on the basis of the raw appearance frequencies. Although most traditional computational approaches begin with the extraction of first moments, and only then the extraction of conditional statistics derived by "normalization," such a sequential process might not apply to biological systems. As previously mentioned in Experiment 3, the information contained in conditional probabilities has ecological advantages. Thus, conditional probabilities may not be derived from joint probabilities, but rather continuously updated as each new element appears in the input stream. The biological apparatus supporting such an on-line computational mechanism would involve a multiplicative interaction or gating between and within neurons, which has been shown to be ubiquitous in the brain (Hausser, Spruston, & Stuart, 2000).

Although our results are not the first to demonstrate that information about shape sequences can be learned, they differ from previous experiments in that they were obtained with precise control over the available spatial and temporal statistics of the sequences, thereby allowing inferences about which specific statistics were learned. Previous visual search studies (e.g., Chun & Jiang, 1998, 1999) have shown that the spatial or featural context in which an object is presented—that is, a joint probability statistic—can facilitate object detection. However, these search tasks involve a target object to which the participant's attention is explicitly directed. Extraction of shape information from images of a rotating object has been claimed to be facilitated by observing the smooth temporal sequence from multiple viewpoints (Stone, 1999), but other results have failed to show such an advantage (Harman & Humphrey, 1999). These discrepancies may have resulted from the fact that these studies used a Type C task (see Figure 1) in which the specific statistics available in the image sequences were not well controlled, thereby allowing participants to use a variety of low-level spatial correlations during learning. In a Type D task, Olson and Chun (2001) showed that attention directed to a specific target shape facilitated reaction times when the preceding context of temporally ordered shapes was predictable. Although predictability minimally involves joint probabilities, their design did not allow for an interpretation based on conditional probabilities.

The results of our experiments are also qualitatively different from the extraction of temporal correlations by low- and mid-level visual analyzers, because the time course of learning is on the order of minutes rather than the immediate (less than 500 ms) latency to detect coherent motion or texture segregation. However, like these lower-level mechanisms, the process of extracting temporal correlations on our tasks was unsupervised. Participants were not instructed to segment or group the shape sequence but were merely told to attend to the display, in contrast to studies of perceptual learning in which a well-defined task and a long learning phase are required for participants to show improvements in performance (Goldstone, 1998). Nevertheless, the fact that participants extracted statistical descriptors as complex as the temporal conditional probability of shape pairs without any instruction supports the hypothesis that this type of unsupervised learning is a common component of the different classes of temporal observational learning reported at various levels in the visual system, including low- and mid-level analyzers (e.g., Ball & Sekuler, 1982; Frensch, Buchner, & Lin, 1994; Vidyasagar & Stuart, 1993).

Our experiments also have some similarities with the SRT literature (see Cleeremans, 1993; Cleeremans, Destrebecqz, & Boyer, 1998). Those studies, like the present experiments, use a continuous stream of visual events, but they also require a motor response as the participants perform a discrete button-pressing task. In SRT tasks, learning can be based on either the temporal statistics of the visual stimuli or on the spatiotemporal characteristics of the motor responses. Thus, care must be taken to explore sensitivity to the temporal statistics embedded in the visual sequences rather than in the motor responses (Koch & Hoffman, 2000; Willingham, 1998). Although many studies have shown that performance improves in an SRT task (e.g., Cohen et al., 1990; Lewicki, Hill, & Bizot, 1988), most of these studies did not control the first-order statistics of movement frequencies (Reed & Johnson, 1994). Studies that have controlled for first-order statistics report that learning is correlated with the level of statistical structure in the sequences (Reed & Johnson, 1994; Stadler, 1992), and Hunt and Aslin (2001) showed that higher-order statistics *alone* can lead to faster reaction times. Although Howard, Mutter, and Howard (1992) showed that observation of the visual stimulus locations in an SRT display (in the absence of the motor responses) transferred to the SRT task, the sequences they used did not differentiate between joint and conditional probabilities.

Our experiments also have some similarities to the literature on artificial grammar learning, in that conditional probabilities for shape pairs could be described by the rules of a finite-state grammar (Reber 1967, 1989). One important difference is that those studies present training sentences with an already segmented input stream (either spatially or temporally), thereby providing the learner with anchor points from which initial statistical computations may proceed. In our experiments, no anchor points were

[3] To definitively answer the question of whether participants learned triplets per se rather than only pair-based statistics, one would need to construct a test where triplet probabilities are uncorrelated with both first-order (appearance frequency) and second-order (covariations of pairs) statistics. Although the first requirement was fulfilled in our experiment, the second was impossible to fulfill within the framework of triplet testing because there are only two pairs in a triplet, and they are connected by the middle element. To generate two test items with identical second-order but different higher-order statistics, one would need the freedom to vary the position of shape pairs within the test items, which is not possible with triplets.

[4] We thank an anonymous reviewer for directing our attention to this issue.

present until after the familiarization phase, when the participants began the 2IFC posttest.

The present results are conceptually identical to those reported in the auditory domain for temporal sequences (Aslin et al., 1998), in the visual domain for spatial configurations (Fiser & Aslin, 2001), and in the visuomotor domain for an SRT task (Hunt & Aslin, 2001). In all these experiments, participants showed strong, automatic sensitivity to conditional probabilities between events, an aspect of human perception that is crucial for effective associative learning of new features (Atick, 1992; Barlow, 1989). Such similarity between the auditory, visual, and visuomotor modalities suggests that, although spatial and temporal information plays a different role in different modalities, the basic mechanisms by which information is processed are very similar across the spatial and temporal domains. Such a uniform mechanism for deriving higher-order descriptions from sensory input may help to create common representations across independent domains or modalities, thereby simplifying the process of interaction with other mechanisms such as attention.

In summary, we have shown that human observers can extract higher-order temporal statistics from a continuous stream of simple shapes. They perform this statistical learning rapidly, and in an unsupervised manner, suggesting that the underlying mechanism is well suited to the learning of visual events in the natural environment. Particularly impressive is the ability to extract a variety of statistics and retain them in memory for computations that may be needed when lower-order statistics are insufficient to make predictions about event sequences. Despite these demonstrations of rapid and robust learning mechanisms, there must be constraints on statistical learning to prevent the combinatorial explosion problem. Further research will be needed to document the full extent of these constraints.

## References


Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science, 9,* 321–324.

Atick, J. J. (1992). Could information theory provide an ecological theory for sensory processing? *Network: Computation in Neural Systems, 3,* 213–251.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review, 61,* 183–193.

Ball, K., & Sekuler, R. (1982, November). A specific and enduring improvement in visual motion discrimination. *Science, 218,* 697–698.

Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages. In W. Rosenblith (Ed.), *Sensory communication* (pp. 217–234). Cambridge, MA: MIT Press.

Barlow, H. B. (1989). Unsupervised learning. *Neural Computation, 1,* 295–311.

Chubb, C., Econopouly, J., & Landy, M. S. (1994). Histogram contrast analysis and the visual segregation of IID textures. *Journal of the Optical Society of America A, 11,* 2350–2374.

Chun, M. M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology, 36,* 28–71.

Chun, M. M., & Jiang, Y. (1999). Top-down attentional guidance based on implicit learning of visual covariation. *Psychological Science, 10,* 360–365.

Chun, M. M., & Nakayama, K. (2000). On the functional role of implicit visual memory for the adaptive deployment of attention across scenes. *Visual Cognition, 7,* 65–81.

Cleeremans, A. (1993). *Mechanisms of implicit learning: Connectionist models of sequence learning.* Cambridge, MA: MIT Press.

Cleeremans, A., Destrebecqz, A., & Boyer, M. (1998). Implicit learning: News from the front. *Trends in Cognitive Sciences, 2,* 406–416.

Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General, 120,* 235–253.

Cohen, A., Ivry, R. I., & Keele, S. W. (1990). Attention and structure in sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 17–30.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science, 12,* 499–504.

Frensch, P. A., Buchner, A., & Lin, J. (1994). Implicit learning of unique and ambiguous serial transitions in the presence and absence of a distractor task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 567–584.

Gallistel, C. R. (1990). *The organization of learning.* Cambridge, MA: MIT Press.

Goldstone, R. (1998). Perceptual learning. *Annual Review of Psychology, 49,* 585–612.

Harman, K. L., & Humphrey, K. (1999). Encoding "regular" and "random" sequences of views of novel three-dimensional objects. *Perception, 28,* 601–615.

Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a nonhuman primate: Statistical learning in cotton top tamarins. *Cognition, 78,* B53–B64.

Hausser, M., Spruston, N., & Stuart, G. J. (2000, October). Diversity and dynamics of dendritic signaling. *Science, 290,* 739–744.

Henderson, J. M. (1997). Transsaccadic memory and integration during real-world object perception. *Psychological Science, 8,* 51–55.

Horowitz, T. S., & Wolfe, J. M. (1998). Visual search has no memory. *Nature, 394,* 575–577.

Howard, J. H., Mutter, S. A., & Howard, D. V. (1992). Serial pattern learning by event observation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 1029–1039.

Hunt, R., & Aslin, R. N. (2001). Statistical learning in a serial reaction time task: Simultaneous extraction of multiple statistics. *Journal of Experimental Psychology: General, 130,* 658–680.

Irwin, D. E. (1996). Integrating information across saccadic eye movements. *Current Directions in Psychological Science, 5,* 94–100.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics, 14,* 201–211.

Julesz, B. (1981). Textons, the elements of texture perception and their interactions. *Nature, 290,* 91–97.

Kellman, P. J., & Short, K. R. (1987). Development of three-dimensional form perception. *Journal of Experimental Psychology: Human Perception and Performance, 13,* 545–557.

Koch, I., & Hoffman J. (2000). The role of stimulus-based and response-based spatial information in sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 836–882.

Lee, S.-H., & Blake, R. (1999, May). Visual form created solely from temporal structure. *Science, 284,* 1165–1168.

Lewicki, P., Hill, T., & Bizot, E. (1988). Acquisition of procedural knowledge about a pattern of stimuli that cannot be articulated. *Cognitive Psychology, 20,* 24–37.

Maljkovic, V., & Nakayama, K. (1994). Priming of pop-out: I. Role of features. *Memory & Cognition, 22,* 657–672.

Maljkovic, V., & Nakayama, K. (1996). Priming of pop-out: II. The role of position. *Perception & Psychophysics, 58,* 977–991.

Maljkovic, V., & Nakayama, K. (2000). Priming of pop-out: III. A short-term implicit memory system beneficial for rapid target selection. *Visual Cognition, 7,* 571–595.

Olson, I. R., & Chun, M. M. (2001). Temporal contextual cueing of visual

attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1299–1313.

Reber, A. S. (1967) Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior, 6,* 855–863.

Reber, A. S. (1989) Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General, 118,* 219–235.

Reed, J., & Johnson, P. (1994). Assessing implicit learning with indirect tests: Determining what is learned about sequence structure. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 585–594.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II* (pp. 64–99). New York: Appleton-Century-Crofts.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996, December). Statistical learning by 8-month-old infants. *Science, 274,* 1926–1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70,* 27–52.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language, 35,* 606–621.

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science, 8,* 101–105.

Simons, D., & Levin, D. (1997). Change blindness. *Trends in Cognitive Sciences, 1,* 261–267.

Stadler, M. A. (1992). Statistical structure and implicit serial learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 318–327.

Stone, J. V. (1999). Object recognition using spatiotemporal signatures. *Vision Research, 38,* 947–951.

von der Malsburg, C. (1981). *The correlation theory of brain function* (Internal Rep. No. 81-2). Göttingen, Germany: Department of Neurobiology, Max-Planck Institute for Biophysical Chemistry.

Vidyasagar, T. R., & Stuart, G. W. (1993). Perceptual learning in seeing form from motion. *Proceedings of the Royal Society, London (B), 254,* 241–244.

Watamaniuk, S. N. J., & Sekuler, R. (1992). Temporal and spatial integration in dynamic random dot stimuli. *Vision Research, 32,* 2341–2348.

Willingham, D. B. (1998). A neuropsychological theory of motor skill learning. *Psychological Review, 105,* 558–584.

Received August 11, 2000
Revision received October 12, 2001
Accepted October 12, 2001 ■